# Contextualizing Heterogeneous Data for Integration and Inference

**Zachary Pincus and Mark A Musen, MD PhD.**

**Stanford Medical Informatics, Stanford University School of Medicine, Stanford, CA**

*Systems that attempt to integrate and analyze data from multiple data sources are greatly aided by the addition of specific semantic and metadata "context" that explicitly describes what a data value means.*

*In this paper, we describe a systematic approach to constructing models of data and their context. Our approach provides a generic "template" for constructing such models. For each data source, a developer creates a customized model by filling in the template with predefined attributes and value. This approach facilitates model construction and provides consistent syntax and semantics among models created with the template. Systems that can process the template structure and attribute values can reason about any model so described.*

*We used the template to create a detailed knowledge base for syndromic surveillance data integration and analysis. The knowledge base provided support for data integration, translation, and analysis methods.*

## THE NEED FOR CONTEXT IN INTEGRATING HETEROGENEOUS DATA

The goal of our ongoing research into syndromic surveillance is to successfully extract early disease indicators from so-called *non-traditional* health surveillance data, such as over-the-counter pharmaceutical sales, school and workplace absenteeism, and Emergency-911 calls. Monitoring these data may be key to detecting the onset of epidemics as early as possible;[1] unfortunately, such data can be noisy and nonspecific. Our system, called the Biological Spatio-Temporal Outbreak Reasoning Module (BioSTORM), aims to provide early and specific detection of epidemics while avoiding the problems of noise by comparing trends across multiple traditional and non-traditional health data sources.[2]

These aims require the ability to integrate conceptually diverse data and to reason about those data in a consistent manner. Unfortunately, there are no pre-existing standards for reporting non-traditional health data, which are extremely heterogeneous in data content as well as semantics. These problems drove us to pursue a very general approach to data integration across multiple and diverse data sources.

While such problems are relatively acute in the area of syndromic surveillance, they are not new to informatics. Automatic comparison of and reasoning about diverse data from different sources remains an open problem. A chief hurdle to such data integration is the fact that, in different data sources, the same concepts can be represented differently and different concepts can be represented in a superficially similar manner.

A common approach to dealing with this hurdle is to place the data in *context* in order to provide detailed metadata to integration or mediation systems.[3] The context of a piece of data includes its semantics ("To what specific concept does this piece of data refer?"), its syntax ("How is this piece of data structured?"), and other related metadata such as information about the quality of the data. Without context, a piece of data is near-meaningless. For example, to understand the fragment of an HL7 message "|*234-7120*~|," one must know both its syntax and semantics. The syntax indicates that a "~" indicates that the preceding entry is to be repeated. The semantic meaning is needed to know that in that particular field, "234-7120" is a phone number that can be used to contact a person. (The HL7 Reference Information Model seeks to make this sort of semantic information explicit.) Traditionally, however, the design of a database and client information system *implicitly* defines the semantics and to some degree even the syntax of a data set.

Such an *ad hoc* approach is unworkable for systems that attempt to integrate arbitrary data sources or to reason about integrated data. Such system need the context of data to be explicitly specified in order to determine whether two pieces of data may be combined, how they might be combined, and what that combination might mean. Two major approaches to the related problems of defining the context of the data in a single data source and of relating contexts across a number of sources have been pursued in the field of information integration and mediation.

The first approach is to create an explicit *local model* for each data source, which describes the context of data in that source. Local models range from database schema that fix the relationships and constraints between data values to more expressive ontologies describing a *domain knowledge base*[4] of the structure and attributes of data from one source. Once such models are created, there are several techniques for bringing together multiple local models, including schema matching,[5] ontology merging and integration,[6] and combined approaches.[7] Unfortunately, constructing local models for many different data sources is labor intensive, and integration methods often do not scale well to dealing with many models.

A second approach to integrating multiple data sources is to design manually a *global model* that specifies the context of an entire domain of knowledge. Specific data sources are then described with explicit reference to this global ontology or schema. This approach has met with much success: SIMS, an early experiment in semantically rich database integration, used a central domain model to tie together multiple databases and to facilitate complex query construction over those databases.[8] More recently, the TAMBIS ontology provides a common umbrella structure that facilitates accessing multiple molecular biology databases by providing a common set of semantics for query formulation.[9] Finally, the caBIO system attempts to provide a set of data objects which both form a model of cancer biology and have built-in methods to transparently query remote databases.[10]

A strict global model is not a panacea, however. If a global ontology is not detailed enough, some data will necessarily be lost to abstraction: For example, if a data source provides records for "city" and "country," but a global model has only "country," representing data in terms of the global model involves the loss of "city" information. Thus, a global model must be both large and detailed in order to accommodate heterogeneous data sources without "abstracting away" potentially relevant information. Further, maintaining a large, detailed ontology is a non-trivial task.

In this paper, we describe a hybrid approach that combines the semantic rigor of creating a global ontology of all data types and sources with the flexibility and level of detail that comes from devising customized local ontologies for each data source. Unlike some other approaches, ours is not limited to describing data in relational databases, and allows for additional metadata beyond syntax and semantics.

The goals of this work were to provide tools for *rapidly* describing and extremely diverse data in a coherent manner that can facilitate reasoning on that data.

## TEMPLATE-DIRECTED ONTOLOGY DEVELOPMENT

Our general approach is to enable data to be self-descriptive by associating them with a structured context. We have defined a very general and reusable structure for describing contexts, which forms a *template ontology* (Figure 1). The template ontology acts as a meta-model, providing a consistent structure within which detailed descriptions of different data sources and their data can be constructed. While our template remains domain-agnostic, it allows detailed descriptions of individual data sources to be constructed.

For each new data source to be described, a developer fills in the template by choosing specific attributes and attribute values from a predefined taxonomy. This template-directed process allows users to create a customized local model that shares a common structure, space of attributes, and set of possible attribute values with all other models so created. Any system that can process our template ontology and attributes can access enough relevant context to reason successfully about data from sources described therewith.

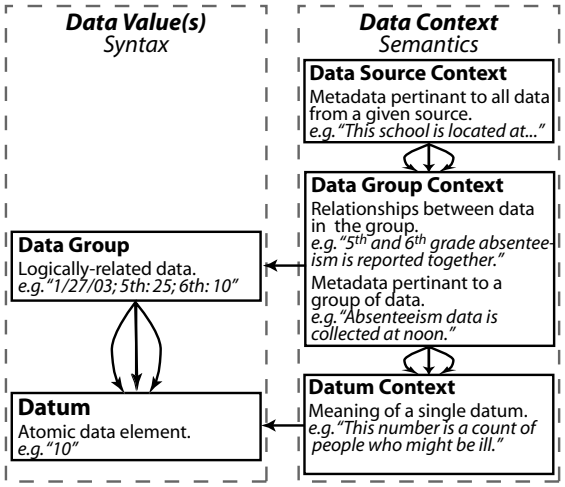


**Figure 1: A Generic Structure for Data and their Metadata Context.** In our template ontology, data values are associated with metadata describing the semantic meaning of the data and absenteeism other relevant context. Arrows indicate one-to-one and one-to-many relationships between concepts.

### The Template Ontology

Our template ontology defines the relevant syntactic and semantic context of a piece of data along several axes (Figure 1). A user describes the context of data from a particular data source by "filling in" the template with relevant details. This requires creating instances of classes from the template ontology at the level of data source, data group, and atomic datum and choosing specific metadata and semantic attributes to fill in the slots of those instances.

The ontology provides a taxonomy of attributes grouped into general categories from which users choose when filling in these slots. For example, the template ontology requires that an `address` value be associated with every `dataSourceContext` instance, but it is up to the user to choose from the provided subclasses of `address`, such as `streetAddress` or `internetAddress`, and create an instance of the chosen address type. Describing a set of data sources is thus reduced to choosing attributes and values from (extensible) taxonomies: the user is neither required nor able to define the structure of the descriptions; they simply fill in the given structure.

Because of the shared structure of the template and its defined vocabulary of possible attributes, it is possible to rationally reason about specific data. For example, a system can use the taxonomic relationships between metadata attributes in the contexts of different pieces of data to data to infer the relationship between those data themselves. (E.g. the fact that

| LOINC Axis | Generic Interpretation | Representation in Template Ontology |
|---|---|---|
| Component/Analyte | What is being measured? *e.g. "Robitussin sales"* | User selects a ìMeasurable Propertyî from a hierarchy of such properties. |
| Kind of Property | How is it being measured? *e.g. "Cases of Robitussin sold per day"* | User selects an attribute from a hierarchy and chooses specific values for that attribute. *e.g. select "rate measure" and choose "per day."* |
| Time Aspect | To what amount of time does a measurement refer? *e.g. "Averaged over a week"* | One datum per data group can be flagged as referring to the time over/at which the group of data was collected. |
| System/Sample | To which region of space does a measurement refer? *e.g. "This pharmacy draws customers from ZIP codes: ... "* | One datum per data group can be flagged as referring to the spatial area/point where the group of data is meaningful. |
| Scale | Is the measurement quantitative, ordinal, nominal, or simply narrative text? | User selects an attribute from a hierarchy and chooses or enters specific values for that attribute. *e.g. select "ordinal" and enter the ordered list of possible measurement values.* |

**Table 1: Generalized LOINC Axes** In our generalization, users choose metadata attributes to "fill in" the five major LOINC axes that provide semantic context for a given atomic datum. Time and space properties apply to every datum in a given group, however: Referring to the same time and place is, in our scheme, necessary for a group of data to be logically related. The metadata attributes exist in modular hierarchies and are user-extendable.

despite outward syntactic differences, two pieces of data both refer to the same semantic concept, "cough syrup sales," can be inferred from explicit contexts.)

### The LOINC Datum Specification

As above, the template ontology provides the structure within which a user "fills in" a description of context at the level of data sources, data groups, and datum elements. We developed such structures for the first two, but to describe individual datum elements we turned to the Logical Identifier Names and Codes (LOINC) terminology.[11] The LOINC scheme, which is used to contextualize results reported by clinical laboratories, does not attempt to enforce a single standard for how data are to be transmitted. Instead, a LOINC specification *describes* what a transmitted datum represents along five major semantic axes. We generalized the LOINC axes from their specific role in reporting clinical lab results to a generic set of descriptors for many different types of reported data (Table 1).

### Domain-Specific Customization

Our template ontology and generalized LOINC specification intentionally provide no domain-specific attributes. To use the template ontology, a user must provide a controlled vocabulary of "Measurable Properties" of relevance to the chosen domain, to which the LOINC descriptions of a datum can refer (see Figure 2, right). Further, a user can optionally add to the taxonomy of pre-defined generic attributes available to describe contexts at the data group and data source levels.

Another kind of customization that our template allows is the creation of new subclasses of the generic data and context classes provided by the template ontology (see Figure 2, left). A user can define subclasses of the data source or data group classes to fit specific requirements. For example, a user could as create a `HospitalContext` subclass of `DataSourceContext` that requires a specific (possibly user-added) metadata attribute such as `numberOfBeds` to be associated with it.

## PROVIDING CONTEXT FOR SYNDROMIC SURVEILLANCE DATA

To evaluate whether this template-directed approach met our goals of facilitating rapid data description and reasoning about described data, we attempted to use our generic ontology to provide context for the non-traditional health surveillance data collected for the BioSTORM syndromic surveillance project. We were able to capture the complexity of the data sources available easily, after simple extensions to the generic ontology and the LOINC "Measurable Property" list (see below). This customized template and the knowledge base describing individual data supported several data integration and analysis methods, each with its own needs for metadata and semantic context.

### The BioSTORM Knowledge Base

We expressed the template ontology in the Protege-2000 ontology development environment,[12] customized it for the "syndromic surveillance" domain, and added descriptions of the data available to the analytic methods developed for the BioSTORM project.

Some of our specific additions to the structure of the template ontology are highlighted in figure 2. We first added a detailed vocabulary of "Measurable Properties" relevant to syndromic surveillance. Next, we added dataSourceContext subclasses for each of the major data sources that might be monitored by a syndromic surveillance system, and created specific metadata attributes for those data sources.

We also added new subclasses of datum to allow BioSTORM to deal with simple time and space properties as atomic entities, and new datum and DataGroupContext subclasses for the space–time data. These provide a vocabulary of attributes that contextualize such data. Finally, we added metadata attributes to allow datum instances to act as "pointers" into a remote database instead of actually containing data values.


**Figure 2: The Template Ontology Customized for Syndromic Surveillance** Additions to the template ontology specific to syndromic surveillance are hilighted. At left is the structure of the template with our added context classes. At right are the top levels of the taxonomy of metadata attributes used to build `LOINCContext` objects. The vocabulary of "Measurable Properties" (shown partially expanded) was our primary addition.

All of these additions were modular and directed by the overall structure of the template ontology. After our customizations, we were able to express all of the recorded data in terms of the LOINC formalism described above, except for some space–time data that used our custom classes. This includes San Francisco emergency-911 call records and detailed patient records from the Palo Alto Veterans Affairs medical center covering demographic information, prescriptions given, tests and procedures ordered, and vital signs recorded.

The customized template was filled in to describe the VA and E-911 data over the course of two days. Using a template made filling in these descriptions quite easy. Further, the template ensured that the descriptions for these two very different data sources remained structurally similar.

**Uses of the BioSTORM Knowledge Base**

To date, the BioSTORM knowledge base has supported three different types of data integration and analysis (Figure 3).

First, the knowledge base supports data retrieval and grouping by data broker software. The data broker uses metadata from the datum classes to retrieve sets of data from a relational database, flat file, or other storage and then formats and groups the data as specified by the context instances (which may be quite different from the original structure and groupings). Finally, the retrieved data values are packaged with the appropriate contexts to create semantically-meaningful data objects. Here, our knowledge base supports data integration by specifying how data is retrieved from distributed databases and by how meaningful data values should be reformatted.

Next, data mapping methods transform data and context between the format of the BioSTORM ontology and of various *input/output ontologies*[13] for analyses performed by generic methods or ones that do not need the full complexity of the BioSTORM knowledge base. Here, off-the-shelf ontology translation software was able to use the knowledge base directly. Data context information from the knowledge base was used to inform the mapper which mappings were valid and meaningful to apply.

The data broker and mapper together act as *data mediators* in the terminology of Wiederhold and others.[14] Mediators sit between distributed databases and end users of such databases and "add value" to the query and data traffic. Here, the value added is the addition of context information by the data broker and the use of that context information to suitably transform the incoming data to more useful formats for analysis. The analytic methods, which are the end consumers of this data, can also make use of the context of the incoming data to perform more inferences about them.

## DISCUSSION

Our initial evaluation suggests that the template-directed approach met our goals. The ontology and fixed set of metadata attributes provided a constrained environment where there was little ambiguity about how data sources should be represented. Compared to the unconstrained task of creating a global model of an entire domain of knowledge from scratch, using a template to describe data sources allowed easy and rapid modeling of heterogeneous data sources.

Further, the additions to the template required for modeling the surveillance data sources were simple, and their scope was limited by the template structure. Our most extensive customization was a controlled vocabulary of "Measurable Properties" for Syndromic Surveillance, with far less complexity than even a basic global model of that domain would require.

Note that the expressivity and computational tractability of a representation directly trade off: the more complex a description can be, the harder it is to use. The fact that our template is restricted allows simple and generic methods to perform inference on it, dem-

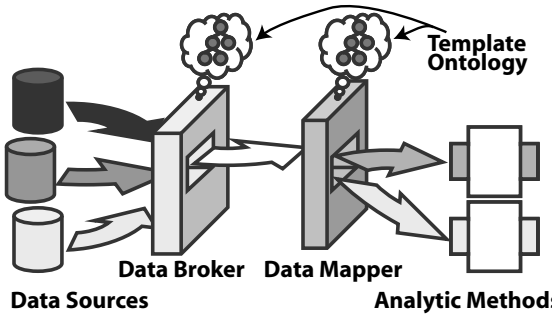


**Figure 3: Providing Context for Syndromic Surveillance.** Our template ontology describes the context of heterogeneous data and data sources. This context supports retrieving data and placing them in a consistent format (Data Broker) and transforming those data (Data Mapper) into new formats suitable for generic analytic methods.

onstrated by our ability to directly apply generic off-the-shelf ontology translation methods.

However, using a restricted template and simple attribute building blocks did not greatly limit the utility of the models produced. Our experiments with the data broker demonstrated that, like SIMS,[7] TAMBIS,[8] and caBIO,[9] which use more complex, free-form global models, the template ontology can support the collection, integration, and analysis of data from remote sources. Currently, it does not support interactive database query methods like those of SIMS and TAMBIS, which use their rich models to construct detailed queries in response to user requests. We do not believe that this limitation is due to the restricted complexity of the template, and providing such an infrastructure is a current research goal. At present, our query model is much more like that of caBIO, which provides objects with internal methods for data retrieval based on pointers to underlying databases.

The limited complexity and regular structure of our underlying model allows for easy development of analytic methods. In contrast, caBIO represents the domain of cancer biology as a set of Java instances, each with a different set of behaviors and properties. SIMS and TAMBIS use formal description logic models that nevertheless have much less structural regularity than our template. Thus, interacting with different elements of these models requires software to be customized for each different element.

In contrast, our template has few major features, which are used regularly in all data source models. As such, creating software to reason about the template is simpler. Further, the hierarchical structure of the template allowed us to construct methods piecewise: Initially the methods used the knowledge base superficially; later we extended the depth of knowledge base use as necessary. Finally, we believe that these methods can be reused with models of different kinds of data sources, because of the simple template structure around which the methods are designed.

In sum, using a template-directed approach to describe multiple data sources produced an infrastructure that supported methods similar to other advanced database integration systems. The template-directed approach also eased the task of modeling data sources and provided a consistent structure around which custom software was easily built and to which generic software were easily applied.

This work was supported by the Defense Advanced Research Projects Agency. Many thanks to Natasha Noy, Monica Crubezy, David Buckeridge, Russ Cucina, Martin O'Connor and Michael Choy for their time and input.

## REFERENCES

1. Proctor ME, Blair KA, Davis JP. Surveillance data for waterborne illness detection: an assessment following a massive waterborne outbreak of Cryptosporidium infection. *Epidemiology & Infection* 1998;120(1):43-54.
2. Buckeridge DL, Graham JK, O'Connor MJ, Choy MK, Tu SW, Musen MA. Knowledge-based bioterrorism surveillance. *Proc AMIA Symp* 2002:76-80.
3. Sciore E, Siegel M, Rosenthal A. Using semantic values to facilitate interoperability among heterogeneous information systems. *ACM Transactions on Database Systems* 1994;19(2):254-90.
4. Gruber TR. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies* 1995;43(5-6):907-28.
5. Rahm E, Bernstein PA. A survey of approaches to automatic schema matching. *VLDB Journal* 2001;10(4):334-50.
6. Noy NF, Musen MA. PromptDiff: A fixed-point agorithm for comparing ontology versions. *The Eighteenth National Conference on Artificial Intelligence (AAAI-02);* August, 2002; Edmonton, Canada.
7. Cali A, Calvanese D, De Giacomo G, Lenzerini M. Accessing data integration systems through conceptual schemas. *10th Italian Conference on Database Systems (SEBD2002)*; June 2002; Portoferraio, Italy.
8. Arens Y. SIMS: Addressing the problem of heterogeneity in databases. *Proc. SPIE* 1997;2938:54-64.
9. Goble CA, Stevens R, Ng G, Bechhofer S, Paton NW, Baker PG, et al. Transparent access to multiple bioinformatics information sources. *IBM Systems Journal* 2001;40(2):532-51.
10. National Cancer Institute. caCORE Technical Guide: ftp://ftp1.nci.nih.gov/pub/cacore/caCORE1.0_Tech_Guide.pdf.
11. McDonald C, Huff SM, Suico J, Mercer K. Logical Observation Identifier Names and Codes Users' Guide. Regenstrief Institute; 2002. http://www.loinc.org/download/loinc/guide/LOINCManual.pdf
12. Gennari JH, Musen MA, Fergerson RW, Grosso WE, M C, Eriksson H, et al. The evolution of Protege: an environment for knowledge-based systems development. *International Journal of Human-Computer Studies* 2003;58(1):89-123.
13. Fensel D, Eriksson H, Musen MA, Studer R. Conceptual and formal specifications of problem-solving methods. *International Journal of Expert Systems Research and Applications* 1996;9(4):507-32.
14. Weiderhold G. Mediation to deal with heterogeneous data sources. *2nd International Conference on Interoperating Geographic Information Systems (INTEROPP'99);* March 1999; Zurich, Switzerland.